Data Quality Assessment of Pedestrian, Bicyclist, and Motorcycle Crashes in Michigan: Final Report

Patrick Bowman
Carol Flannagan
Andrew Leslie
Helen Spradlin

Sept. 30, 2023

1 Introduction 3
2 Data and Methods 3
2.1 Data Sources 3
2.2 Probabilistic Linkage 5
2.2.1 Background 5
2.2.2 Implementation 6
3 Results 7
3.1 Linkage Rates 7
3.2 Geographic Disparities in Completeness 8
3.3 Comparison of Injury Severity 8
3.4 Comparison of Race 10
4 Discussion and Recommendations 10
4.1 Overview 10
4.2 Linkage Rates 10
4.3 Comparison of Variables 11
4.4 E-Codes 12
4.5 Recommendations 12
5 Acknowledgements 13
6 References 13
7 Appendix A Linkage Rates by County 13

1 Introduction

Road safety is one of the most pressing issues facing national, state, and local governments. Over 42 thousand people died in traffic crashes across the US last year, and in just 2022, there were 1,123 traffic fatalities on Michigan roads. The ongoing Towards Zero Deaths campaign and other efforts by the Michigan Department of Transportation (MDOT) and the Michigan State Police (MSP) reflect this focus, but in order to address road safety more broadly there is a need for comprehensive and accurate documentation of the problem.

The primary source of crash data are police crash reports, referred to as UD-10 reports in Michigan. These reports provide a wealth of information about crashes occurring on public roads but are limited to what can be determined by the responding officer. Many aspects of crashes are recorded reliably, but some are difficult to record after the crash and others require subjective judgement. Injury severity, which is central to the road safety issue, unfortunately falls into the second category. The KABCO injury scale used by police includes "possible," "suspected minor," and "suspected serious" injuries which may not correspond to injury determinations made by medical professionals. The prevalence of this type of error was well documented in a 2019 review of the literature by Imprialou and Quddu (2019). As such, using other sources of injury data can be useful in evaluating police reports and determining the overall accuracy of the officer's determinations and whether any systemic errors exist.

This report documents a comparison between police crash reports and emergency medical services (EMS) data, which provide a second source of injury information. Given the aforementioned association between misclassification of injuries and person type, the work documented here focused on three types of crash participants with high injury risks: pedestrians, bicyclists, and motorcyclists. Collectively, these participants have outsized injury risk and represent 25.1% of the police reported fatalities and suspected serious injuries in 2022 despite being only 1.2% of the crashing population. Crash report records for these parties were linked to the EMS data using elements present in both datasets and the reported injury severities were compared for matched records. In addition, the EMS records that could not be matched to crash reports were used to examine potential under-reporting and evaluate any geographic or demographic patterns in the mismatch.

2 Data and Methods

2.1 Data Sources

This project made use of police crash report data from 2018 to 2021. These data were extracted from the UD-10 forms and made available to UMTRI by the Criminal Justice Information Center Traffic Crash Reporting Unit of the Michigan State Police. Person records for pedestrians, bicyclists, and motorcyclists were identified and divided into three separate datasets, with 8,395, 5,554, and 13,041 records, respectively.

The EMS data[1] were provided by the State of Michigan Department of Health & Human Services Bureau of Emergency Preparedness, EMS, and Systems of Care for the same time

[1] Data dictionary available here: https://nemsis.org/media/nemsis_states/repository.html?repository=michigan&file=Resources/MI_StateDataSet.xml&at=refs%2Fheads%2Frelease-3.5.0#

period of 2018 to 2021. Since EMS responds to more than just traffic crashes, we asked for data restricted to all events that could be considered motor vehicle crashes. UMTRI reviewed the MIEMSIS Cause of Injury codes (eInjury.01, which conforms to the National EMS Information System universal data standard) and requested only cases with causes believed to be associated with traffic crashes. The data were then narrowed down to subsets containing only pedestrians, bicyclists, and motorcyclists after additional review of the Cause of Injury codes. Note that some Cause of Injury codes may include crashes that would not qualify for a police crash reports (e.g., non-traffic crashes or pedestrian-bicyclist crashes). Finally, patients were sometimes transported by EMS multiple times (e.g., for hospital transfers), so these additional transport events were filtered out of the datasets by UMTRI to avoid duplicates. After these limitations, there were 19,827 records available for matching, breaking down by person type into 6,558 pedestrians, 4,915 bicyclists, and 8,354 motorcyclists.

Both EMS and UD-10 records contained a wide variety of elements, but the project-relevant variables contained in both datasets were date and time of event; location of event; person age; person gender; person race; degree of injury; and transporting ambulance code. Since these data were collected by different agencies, there were some notable differences in format and content is summarized in Table 1. Two of these differences are of particular note:

1. Race was only added to the UD-10 form in 2021 but was available in all years of the EMS data. As such, comparisons were restricted to 2021 when considering race.
2. Police reports and EMS records have different means of recording injury severity. On the UD-10s, police officers use the KABCO scale, a five-level scale ranging from uninjured to fatally injured shown in the left column of Table 2. The EMS records instead use the revised trauma score (RTS), which ranges from 0 to 12. Since the purpose of this project was to use EMS data to validate the police report characterization of injury, the RTS scores were mapped to KABCO as shown in Table 2.

After extracting the project-relevant variables and standardizing the data as needed, records from the two sources were linked using a probabilistic linkage approach.

Table 1 Differences in project-relevant variables by data source

Variable	UD-10 Police Reports	EMS Records
Event Date and Time	Datetime of crash	Datetime of transport
Event Location	County of crash	County where patient was collected
Person Age	Years of age	Variable, converted days/months to years by UMTRI
Person Gender	Male or Female	Six available codes but only Male and Female were used
Person Race	2021 only, enumerated list	Enumerated list, but multiple codes can be selected
Degree of Injury	KABCO scale	Revised trauma score
Ambulance Code	Free text, optional	Free text, optional

Table 2 Comparison of injury severity scales

KABCO Injury Severity	Revised Trauma Score
Fatal (K)	0-2
Suspected Serious (A)	3-10
Suspected Minor (B), Possible (C), or No Injury (O)	11-12

2.2 Probabilistic Linkage

2.2.1 Background

Linkage is the process of identifying and connecting records in two datasets where the records represent the same individual or event (Sayers et al., 2016). In an ideal scenario, all identifying variables would be present in both datasets with identical formatting, so records could be matched by finding any cases where the identifiers align. This is rarely the case in practice, as differences in how data are recorded from dataset to dataset preclude simple matching. For instance, as shown in Table 1, while the UD-10 records and the EMS records both have an event date and time, the referenced event is different: the crash on the UD-10 and the transport for the EMS. This means that matching records may not have the same time (or even the same date) in the two systems.

To address this, one could encode a system of matching rules for each variable and then assign matches based on the outcome of those rules. This approach, called deterministic linkage, can function well in some circumstances but its utility is limited by the need to define the matching rules. Rules that are too permissive or too strict will produce poor quality matches and rare types of disagreements between the datasets may not be captured by the rules. The simplest version of deterministic linkage links two records on the same given unique identification variable (e.g., social security number) in two datasets. This common identification variable is not present in the crash and EMS datasets. As such, a common approach is to not define concrete rules, but instead look at the general similarity between the records.

This approach, called "probabilistic linkage" compares each of the identifying variables to see how similar they are. For instance, the crash date "January 1, 2022" and the EMS event date "January 2, 2022" are more similar than the crash date "January 1, 2022" and the EMS event date "September 1, 2021" and, thus, may more plausibly refer to the same event. For each variable, such as date, the probability that a matching variable agrees given that the comparison record actually match (m-probability) and the probability that the matching variable agrees given that the pair of records is actually a nonmatch (u-probability) are calculated. After comparing the similarity of all of the identifier variables used for probabilistic linkage, one can estimate the likelihood that two records refer to the same record given their similarity. Match weights are produced for each included variable, based on the ratio of the m- and u-probabilities. The sum of the individual variable match weights produces an overall weight, from which a cutoff point is then determined based on the manual review of the weighted records. Records above the cutoff point are considered to be correct linkages and records below the cutoff point are determined to

be too inadequate to be considered a proper match. The exact calculations of this estimation are omitted here, but there are a few relevant implications discussed below (interested readers may find a more technical discussion in of the methods in Sayers et al., 2016).

Since probabilistic linkage is based the similarity of their identifiers, the commonality of those characteristics can influence how easily the records are linked. For example, if only one crash occurred in Lansing and a dozen occurred in Ann Arbor, it would be a more meaningful indicator of matching if the EMS location is Lansing. This is because there is only one possible crash that EMS case could match, assuming that the city is correctly coded. If the EMS case were instead from Ann Arbor, it could match to any one of the crashes there. This intuition persists in more complex implementations, and probabilistic linkage is more successful in linking uncommon or unique records than very common ones. This can be combatted by including more identifying variables, but since the variables must be present in both datasets and of sufficient quality, there are practical limits on how many variables are available to support the linkage.

A final aspect of linkage that should be discussed is blocking. A blocking variable is an identifier that is not necessarily included in the similarity calculations but instead used to divide the datasets into subsets and only allow matches within the subsets. This is useful behavior because there may be characteristics on which the records must match in order to be linked. For instance, all crashes occurred in Michigan, so if EMS transferred someone from a crash scene in Ohio, it should not be allowed to link to any available crashes. Blocking variables are used both to improve the quality of the linkage and reduce the computational complexity by limiting the potential matches.

2.2.2 Implementation

For this project, linkage was performed using the Registry Plus, Link Plus software[2] developed by the Division of Cancer Prevention and Control at the Centers for Disease Control and Prevention (CDC) and was run separately for pedestrians, bicyclists, and motorcyclists. Each data subset of pedestrians, bicyclists, and motorcyclists for both crash and EMS was imported into the program separately with event date; person age; ambulance identification number; county and gender used as identifier variables and county, month and year used as blocking variables. Typically, personal identifiers, such as name or social security number, are best to use, but identifying information was not provided under the terms of use for the EMS data. Additional variables were tested in the linkage models, but these lowered the overall linkage rate for each subset. A cutoff value weight was used to maximize the number of linkages while still maintaining correct linkage data. This maximizes the number of true positives and minimizes the number of false negatives. The variables used in this probabilistic linkage and their roles in the process are summarized in Table 3.

[2] https://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm

Table 3 Variables used in probabilistic linkage process

Variable	Role	Notes
County	Block	Value-specific
Month	Block	Value-specific
Year	Block	Value-specific
Date	Identifier	MM/DD/YYYY date format
Age (years)	Identifier	Value-specific
Ambulance ID	Identifier	Value-specific
County	Identifier	Value-specific
Gender	Identifier	Value-specific

3 Results

3.1 Linkage Rates

The linkage rates for the EMS data are summarized in Table 4. The rate of successful linage was highest for motorcyclists with 78.08% of EMS records matching to a crash record. The match rate was substantially lower for pedestrian and bicyclist cases (59.93% and 49.66%, respectively). The lower match rate for the non-motorist cases may be due, at least in part, to underreporting of these crashes to police, meaning that some number of the unlinked EMS records may not have an associated police report at all. However, match failure does not guarantee that a case was not reported by the police.

For example, there is some ambiguity in the e-codes provided by EMS, so the linkage rates could be improved by clearer use. For instance, approximately half of the linked bicyclist EMS records had e-codes indicating that they were in nontraffic crashes. Given that these records matched to traffic-crash police reports in high volume, it is not clear whether this is a coding error on the EMS report, a misclassification on the police report, or a frequent error in the linkage. Given the required matching criteria, the last option seems unlikely, so it is likely that EMS and police training leads to different evaluation of whether a crash counts as in-traffic. Thus, the 50% linkage rate is likely to be a combination of EMS cases that were not actually police-reportable, cases that should have been police-reported but weren't, and cases that should have linked but didn't.

Table 4 Summary of linkage results.

Person Type	Available EMS Records	Linked EMS Records	Linkage Rate
Pedestrian	6,558	3,930	59.93%
Bicyclist	4,915	2,441	49.66%
Motorcyclist	8,354	6,523	78.08%

3.2 Geographic Disparities in Completeness

The linkage rates discussed above are for the state of Michigan overall. Given the potential role of underreporting in these rates, it is possible that the linkage rates are not uniform across the state due to a variety of reasons including demographics and police presence. To investigate this, linkage rates were calculated by county. The complete table of linkage rates by county is provided in Appendix A. Compared to the statewide linkage rate of 59.9% for pedestrians, county rates ranged from 48.4% in Newaygo County (15/31 records) to 83.8% in Washtenaw County (275/328 records) for counties with more than 30 pedestrians involved in crashes. The Lower Peninsula linkage rate was 60.9% and the Upper Peninsula linkage rate was 39.8%. For bicyclists, county rates ranged from 9.1% in Ionia County (3/33 records) to 98.3% in Mackinac County (171/174 records) for counties with more than 30 bicyclists involved in crashes. The Lower Peninsula linkage rate was 48.4% and the Upper Peninsula linkage rate was 66.5%. The motorcyclist county rates were much higher overall and ranged from 71.1% in Kalamazoo County (145/204 records) to 100% in Iosco County (30/30 records) for counties with more than 30 motorcyclists involved in crashes. The Lower Peninsula linkage rate was 78.5% and the Upper Peninsula linkage rate was 86.7%.

In general, the geographic pattern of linkage rates was not consistent across the three road user types. Moreover, there was no clear pattern of under- or over-reporting by rural vs. urban counties or lower vs. upper peninsula.

3.3 Comparison of Injury Severity

Table 5, Table 6, and Table 7 show the level of agreement between KABCO and Revised Trauma Score (RTS) for linked pedestrians, bicyclists, and motorcyclists, respectively. In each table, rows indicate cases where the police report indicated K, A, or BCO (combined). The columns indicate whether the corresponding RTS score was lower, equivalent, or higher than the police-reported value. For the 608 matched pedestrian records with a Revised Trauma Score (RTS) reported in the EMS data (Table 5), 62.5% had police-reported injury outcomes consistent with the EMS data. Of the misreported injuries, severity was overestimated in 29.0% and underestimated in 8.6%. There were 3,322 pedestrians with no revised trauma score or uncoded KABCO values. For the 669 linked bicyclists with revised score data (Table 6), 68.6% had consistent police-reported injury outcomes, 13.9% had severity overestimated, and 17.5% had severity underestimated. A total of 1,772 bicyclists were missing revised trauma score or KABCO data. For motorcyclists (Table 7), there were 1,991 with a revised trauma score and KABCO data available. There was a rate of 62.4% for the equivalent revised trauma score to KABCO. Severity was overestimated for 33.6% and underestimated for 4.1%. A total of 4,532 records were missing the revised trauma score or KABCO.

Table 5 Agreement between police-reported and EMS-reported injury severity for linked pedestrian records

Police-Reported Injury Severity	Lower EMS-Reported Injury Severity	Equivalent EMS-Reported Injury Severity	Higher EMS-Reported Injury Severity
Killed (K)	23	9	N/A
Suspected Serious (A)	153	20	5
Suspected Minor/ Possible/No Injury (BCO)	N/A	351	47
Overall	**176 (28.95%)**	**380 (62.50%)**	**52 (8.55%)**

Table 6 Agreement between police-reported and EMS-reported injury severity for linked bicyclist records

Police-Reported Injury Severity	Lower EMS-Reported Injury Severity	Equivalent EMS-Reported Injury Severity	Higher EMS-Reported Injury Severity
Killed (K)	13	5	N/A
Suspected Serious (A)	80	77	2
Suspected Minor/ Possible/No Injury (BCO)	N/A	377	115
Overall	**93 (13.90%)**	**459 (68.61%)**	**117 (17.49%)**

Table 7 Agreement between police-reported and EMS-reported injury severity for linked motorcyclist records

Police-Reported Injury Severity	Lower EMS-Reported Injury Severity	Equivalent EMS-Reported Injury Severity	Higher EMS-Reported Injury Severity
Killed (K)	53	26	N/A
Suspected Serious (A)	615	63	10
Suspected Minor/ Possible/No Injury (BCO)	N/A	1,153	71
Overall	**668 (33.55%)**	**1,242 (62.38%)**	**81 (4.07%)**

Based on the correspondence patterns, EMS reported injury severity is frequently lower than the police estimate, with 29.0% of linked records showing a higher severity on the police report. This is particularly true for suspected serious injuries (A-injuries) which are overwhelmingly marked as lower severity by EMS, with 82.7% of all linked "suspected serious injury" records having a lower severity (i.e., RTS score of 11 or 12). The number of "killed" occupants for which RTS is less severe is surprising, but given that police reports are updated to indicate deaths in the following month and EMS reports are not, this is a plausible pattern.

3.4 Comparison of Race

The other key variable that is present in both the EMS and police-report data (but not used for linkage) is race. This variable is only present in the 2021 crash data, so the comparison was only made for that year. Before comparing values, the race codes were aligned between EMS and crash data to the following: American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, and White.

After standardization of the EMS race variable, the police and EMS datasets have compatible race codes for 77.1% of pedestrians (610/791), 82.3% of bicyclists (372/452), 88.4% of motorcyclists (1,241/1,404) where the records linked and both provided a race code. Mismatches were not biased towards any particular race code.

4 Discussion and Recommendations

4.1 Overview

Data linkage is a way to combine information across databases, enhancing both component databases and subsequent analyses. However, linkage is also used to assess the quality and character of a database by obtaining an external source of information that can be used to evaluate data elements within the database being assessed. In this report, we describe an effort to assess certain elements of the Michigan police-reported crash database using information from the Michigan EMS database. Specifically, we looked at pedestrians, bicyclists, and motorcyclists involved in crashes.

A linkage-based assessment focuses primarily on 1) an overall assessment of what the linkage rate means for the completeness of the assessed database; and 2) comparison of variables that the databases have in common (that are not used for linkage itself). In the latter case, the overlapping variables were injury severity and race.

4.2 Linkage Rates

In this analysis, linkage was done by first selecting relevant EMS cases and then linking to crash cases. Since EMS transport indicates that a crash should be police-reportable, in theory, all EMS records should link to crash records. However, especially with pedestrians and bicyclists, the EMS e-codes that indicate how the injury occurred do not necessarily isolate only police-reportable cases. For example, non-traffic crashes may still be included in the EMS data, even though there is a separate code for non-traffic crashes (for motorcyclists and bicyclists, but not pedestrians).

In addition, the linkage process works best when data are complete and accurate and when cases are fairly unique. In principle, vulnerable road user (VRU) crashes such as the ones studied here should be rare enough to be unique (pedestrians of the same age and gender will not often be hit

in the same location at the same time). However, if data elements are missing, either in the EMS data or the crash data, the process can be less effective.

Thus, the linkage rates cannot be interpreted as estimating the level of underreporting. Instead, linkage rates give an indication of where underreporting might be a problem, and in this case, for which VRU categories. For example, the overall linkage rate for motorcyclists (78%) was much higher than for pedestrians (60%) or bicyclists (50%). This may arise partly because there is less ambiguity in the EMS data about which motorcycle crashes are non-traffic crashes (i.e., the EMS data may have a more "pure" set of police-reportable cases in the e-codes). However, it is likely that these cases, which involve motor vehicles on public roads, are less often underreported than pedestrian and cyclist crashes. Pedestrian underreporting is a particular problem because the overall numbers of cases is higher and the linkage rate is lower.

Geographically, it is not clear that there is a strong pattern of underreporting in certain kinds of areas (e.g., rural vs. urban). Instead, underreporting seems to be worse in some counties than others. Linkage rates are unstable for very small samples, so we only consider rates for counties with at least 30 observations in each category. Using this cutoff, the distribution of linkage rates for pedestrians and motorcycles did not reveal any particular outlier counties. For bicycles, only 3 of 33 EMS bicyclists in Ionia County linked for a rate of 9.1%, well below the next larger rate of 25.6% for Bay County. This suggests that bicycle crashes are underreported in Ionia County (as opposed to the linkage rate being due to data issues such as the EMS data identifying many non-traffic crashes). For counties near the bottom of the link-rate list, it might be worth considering whether underreporting is occurring, even though these are not outliers. For bicycles in particular, three counties (Bay, Leelanau, Lenawee) are linking to fewer than 1/3 of EMS cases.

4.3 Comparison of Variables

The two variables that could be compared in the linked data (for quality assessment) were injury severity and race. The race variable matched in 77% of pedestrians, 82% of bicyclists and 88% of motorcyclists. Mismatches showed no pattern of bias (e.g., always recording "White" if in doubt), suggesting that mismatches might either come from random errors in data entry or differences in how race is identified. For example, it is plausible that in many cases of EMS transport, race must be determined based on the judgment of EMS personnel or the police officer because the patient cannot answer a question. In this case, the different responders may or may not choose the same category. In other cases, the person may be able to state their self-identified race, which would generally lead to better matches. In any case, there is no evidence of significant quality issues with the race variable in the 2021 crash data.

For injury severity, the comparison to EMS is more indirect. EMS and crash data use two different systems for defining injury severity. The Revised Trauma Score (RTS) uses measurement of certain conditions (e.g., blood pressure), so it should be generally repeatable, but it is not a gold-standard measure of injury severity that uses medical diagnosis. RTS has been compared to the Injury Severity Scale (ISS), which is based on medical diagnosis, and they are correlated but do not always agree (Gilpin & Nelson, 1991). In addition, there is no standardized relationship between RTS scores and KABCO. To conduct this analysis, we selected ranges of RTS that should reasonably correspond to K, A, and BCO injury groups.

In spite of the challenges of matching injury scales, 62% of pedestrians and motorcyclists and 69% of bicyclists had injury scores that were in agreement. The more notable feature of the

comparison is that police-reported injury rating for pedestrians and motorcyclists is generally more severe than RTS-based EMS rating. That is, among the 31% of cases that did not match for pedestrians and motorcyclists, the substantial majority were rated as more severe on the police reports. Interestingly, this was not true for bicyclist crashes, where the overall match was higher and errors were equally likely to be rated more or less severe on the police report.

The overestimation of injury severity by police has been observed in previous research (e.g., Farmer, 2003; Flannagan et al., 2013). The current results suggest that police officers may still be overestimating injury severity, particularly in use of the A-injury category.

4.4 E-Codes

One challenge in using EMS data for linkage to crash data is in the utility of e-codes, which define the source of injury, and in this context, the type of road user involved in a crash. Importantly, the original premise of linkage in this project was that any crash case involving EMS transport should result in a crash report. If true, and if linkage itself were very successful, this would mean that linkage failure indicates reporting failure. However, we discovered that the e-codes used by EMS to indicate which cases were pedestrians, bicyclists or motorcyclists were ambiguous or possibly used differently to the point where we cannot be sure whether a given EMS case should have resulted in a police report or even, in some cases, whether an EMS case was a pedestrian, bicyclist, or motorcyclist.

Some e-codes were ambiguous about the patient's role. For example, 121 cases used an e-code defined as "Motor Vehicle Crash, motorcycle vs. pedestrian or animal." With this code, the patient might be a motorcyclist or a pedestrian. Similarly, 654 cases labeled as "Motor Vehicle Crash, vehicle vs pedestrian or animal" can be motor vehicle occupants or pedestrians. In these cases, it is possible to include other types of crash cases in the linkage, but that can affect linkage rate in other ways. In this project, we eliminated ambiguous codes.

Other e-codes indicated that the crash might be a non-traffic crash, but it is not clear that EMS personnel are familiar with what that means in the crash data community. For example, the largest single group of bicyclists (2,426) in the EMS data were coded as "Bicycle / pedal vehicle nontraffic accident." We considered eliminating these, but even with only 33% linkage rate, these cases also produced the largest number of linked cases (799). It is unlikely that all of these are actually non-traffic cases in the police reported data, suggesting that EMS personnel are not using this term in the same way as police officers.

The consequence of this e-code issue is that linkage with EMS might always involve some ambiguity about whether a case in the EMS dataset should be expected to be found in the crash dataset. This, in turn, means that linkage rates are influenced not only by reporting completeness and linkage quality but also by the inclusion of EMS cases. Ideally, e-codes would be used more precisely and effectively by EMS personnel, but this would require some input from the crash-data community.

4.5 Recommendations

The goal of the project was to investigate whether there are data quality issues in VRU crash reports, either in terms of accuracy of variable values or in terms of reporting completeness. The results suggest the following:

- Race (used in 2021) in the police reports was accurate relative to race reported by EMS.

- Police may still be overusing the A-injury category to some degree, though overall accuracy in severity assessment is fairly good (60-70% correspondence with EMS data); further efforts to help officers make more accurate assessments of injury level might improve this, though linkage to medical outcome data is probably the best solution in the long run.
- Linkage rates to EMS data were about 50% for bicyclists, 60% for pedestrians, and 78% for motorcyclists; the geographic pattern of linkage rates did not show any clear patterns, but a few counties may have more significant underreporting of bicycle crashes than expected by chance (Ionia, Bay, Leelanau, Lenawee).

5 Acknowledgements

The authors would like to thank Sabrina Kerr, Johnny Wagner and Kevin Putman of the Michigan Department of Health & Human Services Bureau of Emergency Preparedness, EMS & Systems of Care for their assistance in working through the process and agreement required to obtain EMS data.

The authors would also like to thank Amanda Heinze and Melissa Marinoff from the Michigan State Police Criminal Justice Information Center, Traffic Crash Reporting Unit for their help receiving the Michigan crash data and assistance in using the crash data.

6 References

Farmer, C. M. (2003). Reliability of police-reported information for determining crash and injury severity.

Flannagan, C., Mann, N. C., & Rupp, J. D. (2013). *Measuring serious injuries in traffic crashes* (No. 13-4780).

Gilpin, D. A., & Nelson, P. G. (1991). Revised trauma score: a triage tool in the accident and emergency department. *Injury*, *22*(1), 35-37.

Imprialou, M., & Quddus, M. (2019). Crash data quality for road safety research: current state and futuredirections.Accident Analysis & Prevention,130, 84-90.

Sayers, A., Ben-Shlomo, Y., Blom, A. W., & Steele, F. (2016). Probabilistic record linkage. *International journal of epidemiology*, *45*(3), 954-964.

7 Appendix A Linkage Rates by County

The table below provides the linkage rates by county for each of the three VRU categories. Note that for small samples, rates can become extreme by chance, so the rates in the table should only be considered for counties in which there are a sufficient number of cases to link in the first place. In the text, we use a cutoff of 30.

County	Pedestrian Linkage Rate (%)	Bicyclist Linkage Rate (%)	Motorcyclist Linkage Rate (%)
Alcona	0.0 (N=3)	50.0 (N=2)	100.0 (N=1)
Alger	0.0 (N=1)	(N=0)	92.3 (N=26)
Allegan	33.8 (N=71)	52.1 (N=48)	74.6 (N=126)
Alpena	80.7 (N=31)	20.0 (N=10)	100.0 (N=1)
Antrim	0.0 (N=13)	23.1 (N=13)	85.2 (N=27)
Arenac	100.0 (N=1)	50.0 (N=2)	90.0 (N=10)
Baraga	100.0 (N=2)	50.0 (N=2)	80.0 (N=5)
Barry	58.8 (N=17)	50.0 (N=6)	81.6 (N=49)
Bay	25.6 (N=39)	61.7 (N=47)	74.1 (N=116)
Benzie	35.3 (N=17)	0.0 (N=1)	90.0 (N=10)
Berrien	63.0 (N=27)	60.0 (N=45)	94.1 (N=17)
Branch	54.2 (N=24)	46.4 (N=28)	66.7 (N=6)
Calhoun	66.3 (N=92)	60.2 (N=133)	80.0 (N=15)
Cass	69.2 (N=13)	57.1 (N=14)	85.2 (N=27)
Charlevoix	45.5 (N=33)	70.0 (N=10)	77.8 (N=9)
Cheboygan	15.4 (N=13)	66.7 (N=3)	100.0 (N=22)
Chippewa	55.6 (N=9)	33.3 (N=3)	95.5 (N=22)
Clare	33.3 (N=6)	12.5 (N=8)	80.0 (N=40)
Clinton	54.6 (N=22)	33.3 (N=18)	78.7 (N=47)
Crawford	60.0 (N=5)	25.0 (N=4)	87.5 (N=16)
Delta	31.0 (N=29)	42.1 (N=19)	90.9 (N=11)
Dickinson	66.7 (N=9)	6.7 (N=15)	100.0 (N=5)
Eaton	51.0 (N=51)	57.8 (N=83)	90.4 (N=52)
Emmet	48.1 (N=52)	33.3 (N=9)	88.5 (N=61)
Genesee	52.8 (N=108)	54.3 (N=396)	79.7 (N=473)
Gladwin	15.4 (N=13)	75.0 (N=8)	83.9 (N=31)
Gogebic	0.0 (N=10)	75.0 (N=4)	100.0 (N=8)
Grand Traverse	35.9 (N=103)	56.9 (N=51)	84.0 (N=75)
Gratiot	50.0 (N=14)	50.0 (N=14)	80.0 (N=40)

Hillsdale	42.9 (N=7)	14.3 (N=7)	87.9 (N=33)
Houghton	57.7 (N=26)	40.0 (N=10)	50.0 (N=2)
Huron	55.6 (N=9)	33.3 (N=6)	66.7 (N=12)
Ingham	59.1 (N=232)	69.4 (N=157)	88.2 (N=323)
Ionia	9.1 (N=33)	62.2 (N=45)	73.1 (N=26)
Iosco	50.0 (N=8)	33.3 (N=6)	100.0 (N=30)
Iron	0.0 (N=2)	0.0 (N=1)	(N=0)
Isabella	41.4 (N=29)	66.7 (N=24)	74.6 (N=59)
Jackson	70.8 (N=96)	69.1 (N=97)	72.7 (N=11)
Kalamazoo	54.4 (N=340)	76.9 (N=311)	71.1 (N=204)
Kalkaska	50.0 (N=8)	69.2 (N=13)	72.2 (N=18)
Kent	48.4 (N=337)	67.6 (N=757)	79.4 (N=714)
Keweenaw	0.0 (N=10)	(N=0)	100.0 (N=3)
Lake	30.8 (N=13)	63.6 (N=11)	83.3 (N=6)
Lapeer	62.5 (N=24)	68.8 (N=32)	91.3 (N=80)
Leelanau	28.1 (N=32)	66.7 (N=3)	83.9 (N=31)
Lenawee	31.3 (N=32)	56.8 (N=44)	82.7 (N=75)
Livingston	55.3 (N=85)	72.4 (N=58)	82.9 (N=35)
Luce	50.0 (N=2)	0.0 (N=1)	100.0 (N=10)
Mackinac	98.3 (N=174)	0.0 (N=7)	94.1 (N=17)
Macomb	37.0 (N=216)	54.6 (N=194)	83.2 (N=692)
Manistee	57.1 (N=7)	0.0 (N=7)	89.7 (N=29)
Marquette	46.4 (N=84)	60.0 (N=30)	79.0 (N=19)
Mason	20.0 (N=15)	77.3 (N=22)	62.5 (N=8)
Mecosta	38.5 (N=13)	53.9 (N=13)	90.9 (N=33)
Menominee	14.3 (N=7)	47.6 (N=21)	60.0 (N=25)
Midland	55.6 (N=54)	64.3 (N=28)	93.8 (N=80)
Missaukee	100.0 (N=3)	50.0 (N=2)	88.9 (N=9)
Monroe	57.3 (N=75)	54.5 (N=112)	72.1 (N=43)
Montcalm	52.4 (N=21)	47.1 (N=17)	89.8 (N=49)

Montmorency	0.0 (N=2)	0.0 (N=6)	100.0 (N=10)
Muskegon	54.0 (N=76)	65.1 (N=83)	87.1 (N=186)
Newaygo	60.9 (N=23)	48.4 (N=31)	90.9 (N=22)
Oakland	43.6 (N=429)	55.0 (N=380)	78.6 (N=690)
Oceana	28.6 (N=7)	37.5 (N=8)	89.5 (N=19)
Ogemaw	50.0 (N=2)	0.0 (N=3)	83.3 (N=18)
Ontonagon	12.5 (N=8)	50.0 (N=2)	100.0 (N=10)
Osceola	100.0 (N=10)	31.3 (N=16)	(N=0)
Oscoda	0.0 (N=3)	0.0 (N=2)	75.0 (N=4)
Otsego	87.5 (N=16)	63.6 (N=11)	100.0 (N=13)
Ottawa	44.6 (N=211)	62.9 (N=140)	77.1 (N=218)
Presque Isle	0.0 (N=6)	0.0 (N=2)	100.0 (N=4)
Roscommon	75.0 (N=12)	0.0 (N=5)	87.0 (N=23)
Saginaw	67.3 (N=55)	63.2 (N=106)	81.1 (N=127)
Sanilac	60.0 (N=10)	25.0 (N=12)	75.0 (N=40)
Schoolcraft	0.0 (N=3)	0.0 (N=3)	80.0 (N=10)
Shiawassee	41.4 (N=29)	52.6 (N=19)	75.8 (N=62)
St. Clair	72.3 (N=94)	71.6 (N=109)	81.0 (N=58)
St. Joseph	46.4 (N=28)	64.9 (N=37)	74.1 (N=27)
Tuscola	0.0 (N=7)	44.4 (N=9)	71.6 (N=74)
Van Buren	39.3 (N=56)	54.3 (N=35)	82.4 (N=91)
Washtenaw	60.1 (N=416)	83.8 (N=328)	80.7 (N=93)
Wayne	33.4 (N=515)	56.8 (N=2039)	71.7 (N=2338)
Wexford	37.5 (N=8)	62.5 (N=16)	90.9 (N=22)
Total	49.7 (N=4915)	59.9 (N=6558)	78.1 (N=8354)